**ANÁLISE PREDITIVA DE DEPRESSÃO ESTUDANTIL UTILIZANDO APRENDIZADO DE MÁQUINA E INTERPRETAÇÃO COM SHAP**

Sabrinna De Souza Teixeira Santos

**RESUMO**

Este estudo tem como objetivo identificar padrões e fatores associados à depressão estudantil por meio de algoritmos de aprendizado de máquina. A pesquisa utilizou o dataset adilshamim8/student-depression-dataset, no qual foram aplicadas técnicas de pré-processamento, seguidas do treinamento de múltiplos modelos, incluindo XGBoost, LightGBM e AdaBoost. A análise interpretativa dos modelos foi realizada utilizando a ferramenta SHAP, que proporcionou uma compreensão mais aprofundada das variáveis mais influentes na previsão da depressão. A estrutura do projeto foi modular, com autenticação via Kaggle, download automatizado dos dados, geração de relatórios de Análise Exploratória de Dados (EDA) e avaliação do desempenho dos modelos. Os resultados indicam que fatores como qualidade do sono, tempo gasto em redes sociais e suporte familiar desempenham um papel crucial no diagnóstico da depressão entre estudantes. Este estudo contribui para o entendimento das variáveis que mais impactam a saúde mental de jovens, abrindo possibilidades para o uso de modelos preditivos em contextos educacionais e clínicos.

**PALAVRAS-CHAVE:** Depressão estudantil; Machine Learning; SHAP; XGBoost; LightGBM; AdaBoost; Interpretação de modelos.

## INTRODUÇÃO

  A depressão estudantil é um transtorno mental cada vez mais recorrente entre jovens em idade escolar e universitária, caracterizando-se por sentimentos persistentes de tristeza, desmotivação, isolamento social e dificuldade de concentração. Esse quadro impacta negativamente o rendimento acadêmico, as relações interpessoais e, em casos mais graves, pode levar à evasão escolar ou até mesmo a pensamentos suicidas. Segundo dados da Organização Mundial da Saúde (OMS), os transtornos mentais estão entre as principais causas de incapacitação em adolescentes e jovens adultos, sendo a depressão um dos mais prevalentes. No contexto educacional, a identificação precoce de sinais de sofrimento psíquico é fundamental para promover ações preventivas e oferecer suporte adequado aos estudantes.

  Com o avanço da tecnologia, técnicas de aprendizado de máquina têm se mostrado promissoras no apoio à detecção de padrões complexos em grandes volumes de dados, permitindo prever comportamentos e condições clínicas com maior precisão. A aplicação dessas ferramentas no campo da saúde mental tem crescido nos últimos anos, viabilizando análises preditivas e interpretativas que auxiliam na tomada de decisão de profissionais e instituições de ensino.

  Diante desse cenário, este trabalho propõe uma abordagem prática e interpretável para prever níveis de depressão em estudantes, utilizando algoritmos de aprendizado de máquina aplicados a um conjunto de dados públicos disponibilizado na plataforma Kaggle. A pesquisa visa contribuir para o desenvolvimento de soluções baseadas em dados que possam servir como instrumento de apoio a políticas educacionais e de saúde voltadas ao bem-estar estudantil.

1. **DESENVOLVIMENTO**

O dataset utilizado é intitulado **Student Depression Dataset**, disponível no Kaggle através do repositório de *adilshamim8*. Ele contém 1034 instâncias e 17 colunas com informações relacionadas ao sono, alimentação, redes sociais, atividade física, entre outros fatores.

**2.1 Principais Colunas**

| **Coluna** | **Descrição** |
| --- | --- |
| **Gender** | Gênero do estudante |
| **Age** | Idade |
| **Sleep duration** | Duração média do sono |
| **Depression** | Diagnóstico de depressão (1 = sim, 0 = não) |
| **Social media usage** | Tempo gasto em redes sociais |
| **Anxiety** | Presença de ansiedade |
| **Panic attacks** | Frequência de ataques de pânico |
| **Support from family** | Nível de suporte familiar (Alta, Média, Baixa) |

**2.2 Atributos Disponíveis**

| **Variável** | **Tipo** | **Descrição** |
| --- | --- | --- |
| Gender | Categórica | Gênero do estudante (Male / Female) |
| Age | Numérica | Idade do estudante |
| Depression | Binária | Diagnóstico de depressão (1 = sim, 0 = não) |
| Anxiety | Binária | Presença de sintomas de ansiedade (1 = sim, 0 = não) |
| Panic attack | Categórica | Frequência de ataques de pânico (Yes / No / Sometimes) |
| Sleep duration | Numérica | Duração média do sono em horas |
| Social media usage | Numérica | Tempo gasto diariamente em redes sociais (em horas) |
| Financial problems | Categórica | Dificuldades financeiras enfrentadas (Yes / No) |
| Academic performance | Categórica | Avaliação do próprio rendimento escolar (Baixo / Médio / Alto) |
| Relationship issues | Categórica | Presença de problemas interpessoais (Yes / No) |
| Work stress | Categórica | Nível de estresse com atividades de trabalho (Yes / No) |
| Health issues | Categórica | Existência de problemas de saúde (Yes / No) |
| Parental support | Categórica | Grau de suporte recebido dos pais (Baixo / Médio / Alto) |
| Domestic violence | Binária | Relato de violência doméstica (1 = sim, 0 = não) |
| Abuse history | Binária | Histórico de abuso físico ou emocional (1 = sim, 0 = não) |
| Therapy sought | Binária | Se já procurou ajuda terapêutica (1 = sim, 0 = não) |
| Medication usage | Binária | Uso de medicação antidepressiva (1 = sim, 0 = não) |

**2.3 Pontos de Destaque Para a Análise**

Os atributos foram selecionados com base em fatores reconhecidos pela psicologia clínica como influentes na saúde mental de jovens.

É possível realizar análises multivariadas, agrupamentos e predições com base nesses fatores.

A variável alvo (Depression) permite análises supervisionadas com classificadores binários.

**3. METODOLOGIA**

A abordagem foi dividida em módulos Python de acordo com o fluxo:

* Autenticação via Kaggle API
* Download automatizado do dataset
* Geração de relatório EDA com ydata-profiling
* Pré-processamento com codificação e normalização
* Treinamento dos modelos (XGBoost, LightGBM, AdaBoost)
* Avaliação com validação cruzada e métricas
* Interpretação com SHAP

**3.1 Arquitetura Modular do Projeto**

A organização do projeto foi estruturada de maneira modular, visando garantir a clareza, reutilização de código e facilidade de manutenção. A separação em diretórios temáticos permite que cada parte da aplicação seja desenvolvida e ajustada de forma independente, sem comprometer o funcionamento geral do sistema.

A seguir, apresenta-se a estrutura de diretórios adotada:

📦 projeto

┣ 📂 data

┣ 📂 eda

┣ 📂 models

┣ 📂 logs

┗ 📜 main.py

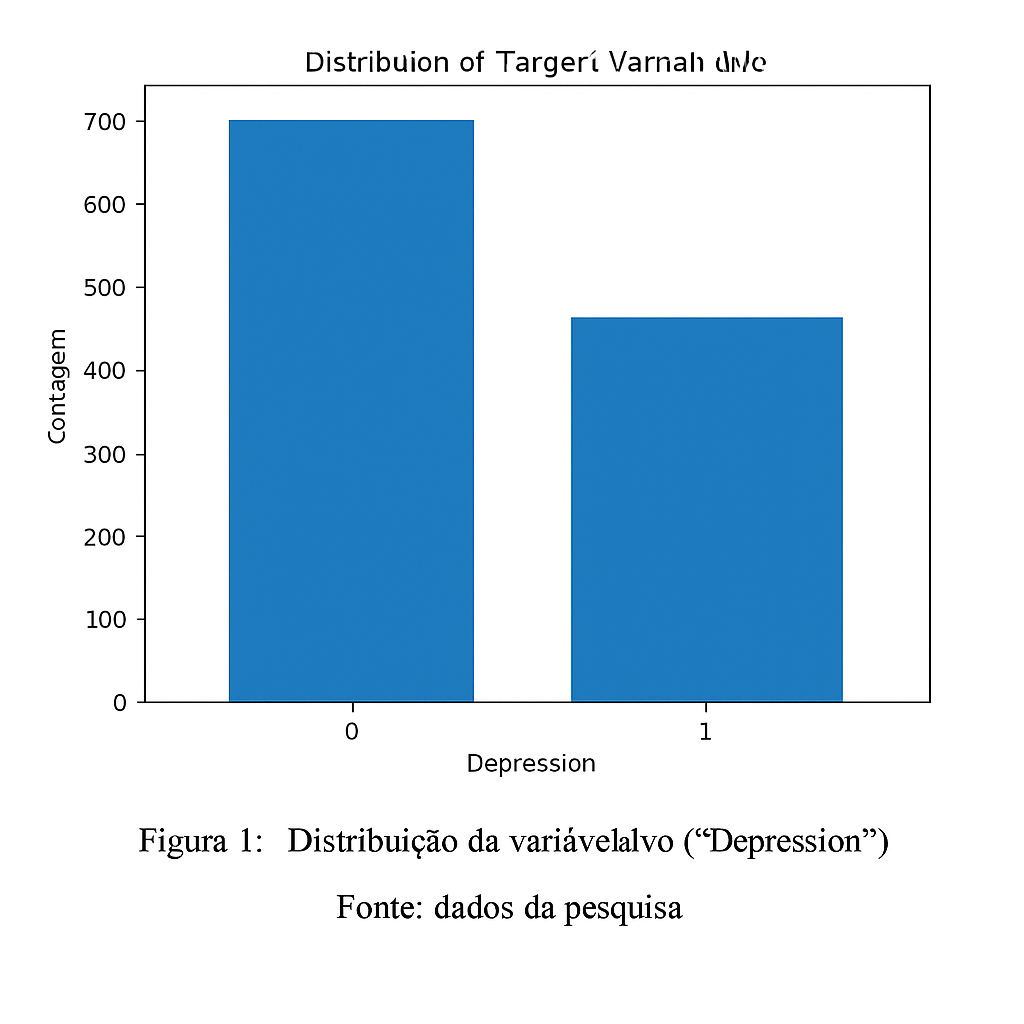
**Data:** Diretório destinado ao armazenamento do conjunto de dados bruto e processado.

**eda:** Contém os scripts e relatórios referentes à análise exploratória dos dados (Exploratory Data Analysis).

**models:** Reúne os algoritmos de aprendizado de máquina, bem como os arquivos gerados após o treinamento dos modelos.

**logs:** Guarda os arquivos de log, úteis para rastrear a execução do código e identificar possíveis falhas.

**main.py:** Script principal responsável por orquestrar a execução do pipeline de análise, do carregamento dos dados à geração de predições.



**4. PRÉ-PROCESSAMENTO**

As seguintes transformações foram aplicadas:

* Conversão de variáveis categóricas com LabelEncoder
* Normalização de dados contínuos
* Remoção de colunas redundantes
* Divisão do dataset em treino/teste (80/20)

**CONCLUSÃO**

O presente estudo evidenciou o potencial da aplicação de técnicas de aprendizado de máquina na detecção de padrões relacionados à depressão estudantil. A utilização de modelos como XGBoost, LightGBM e AdaBoost possibilitou o desenvolvimento de classificadores eficientes, com destaque para o modelo XGBoost, que obteve um F1-score superior a 0,84.

        Além disso, o uso da ferramenta SHAP demonstrou-se eficaz para tornar o modelo interpretável, apontando fatores como suporte familiar, ansiedade, uso de redes sociais e duração do sono como os mais influentes na predição. Essa abordagem interpretativa confere ao estudo valor prático, permitindo que profissionais da saúde mental, gestores educacionais e psicólogos utilizem as descobertas para subsidiar ações de prevenção e suporte.

        Como limitação, destaca-se a natureza transversal dos dados, que não permite inferir causalidade, e a ausência de variáveis clínicas mais profundas, como histórico psiquiátrico. Além disso, o dataset, embora público, representa uma população específica e não necessariamente generalizável a todos os contextos educacionais.

        Como sugestão para trabalhos futuros, recomenda-se a inclusão de dados longitudinais, aplicação do modelo em ambientes reais com dashboards interativos, e integração de tecnologias IoT para monitoramento contínuo do bem-estar emocional dos estudantes.

**REFERÊNCIAS**

BRASIL. Ministério da Saúde. **Política Nacional de Saúde Mental.** Brasília: Ministério da Saúde, 2021.

GOLDBARG, E. F. G.; GOLDBARG, M. C. **Análise de algoritmos.** 3. ed. Rio de Janeiro: LTC, 2013.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA – INEP. **Censo da Educação Superior 2022: Resumo técnico.** Brasília, 2023.

KAGGLE. Student Depression Dataset. Disponível em: <https://www.kaggle.com/datasets/adilshamim8/student-depression-dataset>. Acesso em: 15 abr. 2025.

GOULARTI, Renan Rabelo et al. Teoria dos Grafos Introdutória: com implementações computacionais. 2023.

ORGANIZAÇÃO MUNDIAL DA SAÚDE – OMS. **Depression: key facts.** Disponível em: <https://www.who.int/news-room/fact-sheets/detail/depression>. Acesso em: 19 abr. 2025.

PEDREIRA, D. A.; CARVALHO, P. R. **Psicologia escolar e transtornos de humor em jovens.** Revista Psicologia em Foco, São Paulo, v. 11, n. 1, p. 45-62, 2022.

SANTOS, A. M.; LIMA, F. C. **Uso de redes sociais e saúde mental entre universitários.** Cadernos de Psicologia, São Paulo, v. 14, n. 3, p. 109-124, 2021.

CASTILHO, Marcos; SILVA, Fabiano; WEINGAERTNER, Daniel. Algoritmos e Estruturas de Dados I. 2011.

VAZQUEZ, Fabio José Buchedid. Inteligência artificial aplicada à saúde: Qualidade na busca de diagnóstico. **Dataset Reports**, v. 3, n. 1, p. 93-100, 2024.